Chemometrics and Intelligent Laboratory Systems 170 (2017) 38–50

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

A sparse partial least squares algorithm based on sure independence screening method

Xiangnan Xu [a], Kian-Kai Cheng [b], Lingli Deng [c], Jiyang Dong [a,*]

[a] Department of Electronic Science, Xiamen University, Xiamen 361005, China
[b] Department of Bioprocess & Polymer Engineering and Innovation Centre in Agritechnology, Universiti Teknologi Malaysia, Johor 81310, Malaysia
[c] Department of Information Engineering, East China University of Technology, Nanchang 330013, China

ARTICLE INFO

Keywords:
Sure independence screening
Partial least squares
Variable selection
Asymptotic analyses
Regression analyses
Bayesian information criterion

ABSTRACT

Partial least squares (PLS) regression is a dimension reduction method used in many areas of scientific discoveries. However, it has been shown that the consistency property of the PLS algorithm does not extend to cases with very large number of variables p and small number of samples n (*i.e.*, $p > > n$). To overcome the issue, sparsity can be imposed to the dimension reduction step of the PLS algorithm. This leads to a sparse version of PLS (SPLS) algorithm which can achieve dimension reduction and variable selection simultaneously. Here, we present a new SPLS method called sure-independence-screening based sparse partial least squares (SIS-SPLS) algorithm, by incorporating both SIS method and extended Bayesian information criterion (BIC) into the PLS algorithm. The developed SIS-SPLS method was evaluated using a number of numerical studies including simulation and real datasets. The current results showed that the proposed SIS-SPLS method is efficient in variable selection. It offered low mean squared prediction errors with high sensitivity and specificity. The SIS-SPLS algorithm proposed in the current work may serve as an alternative SPLS method for the analysis of modern biological data.

1. Introduction and motivation

Partial least squares (PLS) regression was introduced by Herman Wold in 1966 [1]. Nowadays, it has been widely used to analyze multivariate data with large number of variables (p) and small sample size (n) (e.g. data generated from –omics experiments). The method models relations between multivariate measurements [2] and reduces the number of variables to a smaller number of latent variables [3]. The PLS algorithm is computationally fast, and it facilitates graphical visualization and interpretation of the original high dimensional data [4].

Despite the attractive properties of the PLS method, researchers had reported some shortcomings of its algorithm. Notably, the standard PLS method does not automatically lead to variable selection [5], and therefore the interpretation of resulting PLS model is not straightforward. The latent variables in the PLS consist of a combination of all original variables, but in most cases only a small portion of the original variables contribute significantly to the projection.

To overcome this issue, researchers had incorporated sparsity into the PLS algorithm to produce sparse PLS (SPLS) methods, which can achieve dimension reduction and variable selection simultaneously. For example, a SPLS method proposed by Chun and Keles [4] imposes sparsity by using an l_1-norm penalty in the procedure of dimension reduction, leading to a sufficient sparse model that enables simultaneous dimension reduction and variable selection. However, the method uses a same tuning parameter for different SPLS components, which may lead to selection of spurious variables in the resulting model. One of the possible solutions for this is to incorporate sure independence screening (SIS) method proposed by Fan and Lv [6]. By using correlation learning to reduce the dimensionality from high to a moderate scale (*i.e.* below sample size), Fan and Lv showed that under some moderate assumptions, SIS can select a model which include all relevant variables with probability asymptotically tends to 1.

The aim of the current study is to integrate the SIS method into the PLS algorithm to obtain a sparse PLS model which preserves the attractive properties of PLS as well as the asymptotic property of the SIS method. This method was named sure-independence-screening based sparse partial least squares (SIS-SPLS). Numerical studies were carried out using both synthetic and real data, and the current results suggested that the developed SIS-SPLS method is efficient and leads to sparse and accurate models.

The rest of this paper is organized as follows. First, we review general principles and properties of the PLS method in Section 2. The SIS-SPLS

* Corresponding author.
E-mail address: jydong@xmu.edu.cn (J. Dong).

https://doi.org/10.1016/j.chemolab.2017.09.011
Received 13 April 2017; Received in revised form 15 August 2017; Accepted 18 September 2017
Available online 21 September 2017

method and its properties are introduced in Section 3. Numerical studies and discussions are provided in Section 4 and 5. In addition, mathematical proofs of key properties of the proposed SIS-SPLS method are given in Appendix.

2. Related works

2.1. Partial least squares

For a given dataset with n samples, q responses and p variables, let response matrix $\mathbf{Y}_{n\times q} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_q)$ and predictor matrix $\mathbf{X}_{n\times p} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$. Without losing generality, assume that $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_q$ are mean-centered, and $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$ are mean-centered and scaled to unit variance.

In the PLS algorithm, latent components $\mathbf{T}_{n\times H} = (\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_H)$ are computed from successive optimization problems. In the first iteration, the loadings on $\mathbf{X}$ and $\mathbf{Y}$ (denoted as $\mathbf{w}_1, \mathbf{c}_1$) satisfies both maximum variances of $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ and $\mathbf{u}_1 = \mathbf{Y}\mathbf{c}_1$, as well as correlation between $\mathbf{t}_1$ and $\mathbf{u}_1$, which can be obtained by:

$$arg \max_{\mathbf{w}_1,\mathbf{c}_1}\left[corr(\mathbf{Y}\mathbf{c}_1, \mathbf{X}\mathbf{w}_1)\sqrt{var(\mathbf{X}\mathbf{w}_1)var(\mathbf{Y}\mathbf{c}_1)}\right] \tag{1}$$

$$s.t. \quad \|\mathbf{w}_1\|_2 = 1, \quad \|\mathbf{c}_1\|_2 = 1$$

where the $\|\bullet\|_2$ is the l_2-norm. The solution of the direction vector (loading) $\mathbf{w}_1$ is the eigenvector of $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$ corresponding to the largest eigenvalue λ_1, *i.e.*

$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}_1 = \lambda_1\mathbf{w}_1. \tag{2}$$

then the score $\mathbf{t}_1$ which is the projection of $\mathbf{X}$ on the direction $\mathbf{w}_1$ can be obtained by:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1. \tag{3}$$

Then, ordinary least squares regression is conducted on $\mathbf{X}$ using $\mathbf{t}_1$, and the residual of $\mathbf{X}$ can then be calculated with Eq. (4):

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T \tag{4}$$

where $\mathbf{p}_1$ is the loading and can be expressed as $\mathbf{p}_1 = \frac{\mathbf{X}^T\mathbf{t}_1}{\|\mathbf{t}_1\|^2}$. By replacing $\mathbf{X}$ with $\mathbf{E}_1$ and $\mathbf{Y} = \mathbf{Y} - \mathbf{t}_1\mathbf{q}_1^T$ in Eq. (1), the coefficient $\mathbf{w}_2$ and score $\mathbf{t}_2$ of second PLS component can be calculated. After h iterations, scores matrix $\mathbf{T}$ with h latent components can be obtained.

For the PLS algorithm, the columns of matrix $\mathbf{T}$ are orthogonal to each other. In addition, columns of matrix $\mathbf{W}$ are also orthogonal to each other. If $\mathbf{Y}$ is univariate [7], and $\mathbf{X}$ and $\mathbf{Y}$ have the relationship $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the $\boldsymbol{\varepsilon}$ is a random vector of normal distribution with the variance $\boldsymbol{\sigma}^2$, then the estimation of $\boldsymbol{\beta}$ using the PLS method can be expressed as

$$\widehat{\boldsymbol{\beta}}^{\text{PLS}} = \mathbf{W}(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{Y} \tag{5}$$

The proof of Eq. (5) is shown in Ref. [8].

Another desirable property of the PLS algorithm is its consistency. For univariate $\mathbf{Y}$, it has been previously shown that under some moderate regularity conditions, $\|\widehat{\boldsymbol{\beta}}^{\text{PLS}} - \boldsymbol{\beta}\| \to 0$ if $p/n \to 0$ in probability, and $\|\widehat{\boldsymbol{\beta}}^{\text{PLS}} - \boldsymbol{\beta}\| > 0$ if $p/n \to k_0 > 0$ in probability [4].

2.2. Sparse partial least squares

Previously, a sparse version of principal component analysis (SPCA) was developed [21]. SPCA achieves sparsity by imposing l_1-norm penalty onto a surrogate of the direction vector ($\mathbf{c}$) instead of the original direction vector ($\mathbf{w}$) while keeping $\mathbf{w}$ and $\mathbf{c}$ close to each other. Later in 2010, SPLS methods [4] were introduced by using the similar strategy. The SPLS methods incorporated the penalty of surrogate direction vector into the PLS algorithm by solving the following optimization,

$$\min_{\mathbf{w},\mathbf{c}}\left[-\kappa\mathbf{w}^T\mathbf{M}\mathbf{w} + (1-\kappa)(\mathbf{c}-\mathbf{w})^T\mathbf{M}(\mathbf{c}-\mathbf{w}) + \lambda_1\|\mathbf{c}_1\| + \lambda_2\|\mathbf{c}_2^2\|\right] \tag{6}$$

$$s.t. \quad \mathbf{w}^T\mathbf{w} = 1$$

where $\mathbf{w}$ and $\mathbf{c}$ are the original direction vector and the surrogate direction vector, respectively. In addition, κ, λ_1 and λ_2 are penalty factors and $\mathbf{M} = \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$.

In Eq. (6), the first term is the object function of the PLS method with a scaling factor κ, and the second term measures the difference between the surrogate direction vector and the original direction vector. The last two terms are the penalty on the surrogate directional vector. This optimization problem can be efficiently solved by the algorithm described in Ref. [4]. For univariate $\mathbf{Y}$, the parameter can be chosen as $\kappa = \frac{1}{2}$, $\lambda_2 \to \infty$, the solution can then be formulated as:

$$\widehat{\mathbf{c}} = \left(|\mathbf{Z}| - \frac{\lambda_1}{2}\right)_+ sign(\mathbf{Z}) \tag{7}$$

where $\mathbf{Z} = \frac{\mathbf{X}^T\mathbf{Y}}{\|\mathbf{X}^T\mathbf{Y}\|}$ and $(x)_+ = max(0, x)$. Chun and Keles recast this soft thresholding as

$$\widehat{\mathbf{c}} = \left(|\mathbf{Z}| - \eta \max_{1\le j\le p}|Z_j|\right)_+ sign(\mathbf{Z}) \tag{8}$$

where $0 \le \eta \le 1$. If $\eta = 0$, SPLS becomes a standard PLS method, and if $\eta = 1$, SPLS gives zeros estimation. Furthermore, a variable will have a higher chance to be selected in the resulting model if it has a higher correlation with $\mathbf{Y}$.

3. SIS-SPLS

3.1. SIS-SPLS algorithm

For univariate $\mathbf{Y}$, the value of $\mathbf{Z}$ in Eq. (8), *i.e.* $\mathbf{Z} = \frac{\mathbf{X}^T\mathbf{Y}}{\|\mathbf{X}^T\mathbf{Y}\|}$), is proportional to the correlation between $\mathbf{X}$ and $\mathbf{Y}$. In the proposed SIS-SPLS method, variables with the correlation larger than $\eta \max_{1\le j\le p}|Z_j|$ will be selected into the resulting model. This step is similar to the SIS procedure proposed by Fan and Lv [6], and can be computed as follows: first, calculate the correlation between $\mathbf{X}$ and $\mathbf{Y}$, then sort the variables based on the correlation, and use first d ($d < p$) variables with largest d correlation to establish a model, and the procedure will be repeated for several times. An interesting property of SIS is that under some regularity conditions, all important variables will be included in the model with probability tend to 1. Instead of using soft thresholding as in SPLS method shown in Eq. (8), the current work used a hard thresholding as in the SIS method. For univariate $\mathbf{Y}$, the performance of PLS algorithm described in section 2 depends on the space spanned by the columns of $\mathbf{W}$, regardless of a scaling factor. One possible option of $\mathbf{w}$ in the h^{th} latent variable is

$$\mathbf{w}^{(h)} = \mathbf{E}_{h-1}^{\text{T}}\mathbf{F}_{h-1} \tag{9}$$

where $\mathbf{E}_{h-1}$, $\mathbf{F}_{h-1}$ are residuals of $h-1$ iteration and $\mathbf{E}_0 = \mathbf{X}$, $\mathbf{F}_0 = \mathbf{Y}$, respectively. Notably, Eq. (9) also has a form of correlation of $\mathbf{E}_{h-1}$ and $\mathbf{F}_{h-1}$ regardless of a scaling factor.

In the proposed algorithm, it is assumed that d_h variables are added in each SIS iteration and the index set of selected variables in h iteration is denoted as A_h. A_h contains the variables of A_{h-1} together with the variables with first d_h largest correlation of current $\mathbf{E}_{h-1}$ with $\mathbf{Y}$. In